I have tested the Qwen2.5-7B-Instruct model very regularly, in which Qwen2.5-7B-Instruct performed well on all parameters.

My recommendation: Qwen2.5-7B-Instruct installed model over Mistral.

Key Strengths:

- **Multi-turn conversation:** It can remember previous messages within the same session and respond appropriately.
- **Supports system instructions:** You can set rules or context at the start of a conversation.
- **Multilingual:** Can handle questions in multiple languages, including Arabic and Hindi.
- **Structured output:** Can produce answers in structured formats like JSON, useful for reporting or downstream applications.
- **Long context handling:** Can process large documents or long conversations, remembering important details.

2. Chat & Context Handling

- Supports chat history and system prompts via Hugging Face `tokenizer.apply_chat_template`.
- Handles multi-turn conversations within a context window of 32k tokens (scalable to 128k with YaRN).
- No built-in persistent memory; session context is maintained only during the active conversation.

4. Function Calling & RAG

- No native function calling; can output JSON describing function calls, which must be handled externally.
- RAG can be implemented by injecting retrieved context into the prompt; not automatic.
- Works best with prompt engineering for domain-specific queries.

Why Qwen2.5-7B-Instruct would be a better choice than Mistral-7B-Instruct-v0.3

Stronger Reasoning & Math Capabilities

- Qwen2.5-7B-Instruct is specifically improved for multi-step reasoning, math, and structured problem-solving.
- Benchmarks like MATH and GSM8K show it outperforms Mistral on complex reasoning tasks.

Better Long-Context Handling

- Default context length is 32k tokens, scalable up to 128k using YaRN.
- This allows it to process long documents, multi-turn chats, or extensive financial data more effectively than Mistral.

Structured Output Support

- Optimized for producing JSON, tables, and other structured formats.
- Makes it ideal for financial reports, automated summaries, or RAG pipelines, whereas Mistral is more text-focused.

Multilingual Proficiency

- Supports 29+ languages, including Arabic, Hindi, and other non-Latin scripts.
- Better suited for global or multilingual deployments, while Mistral has less documented multilingual performance.

Enhanced Instruction Following

- More resilient to diverse system prompts, conditional instructions, and multi-turn chat sessions.
- Maintains context and follows instructions more reliably than Mistral-7B-Instruct-v0.3, especially in specialized domains like finance or coding.

Conclusion

Qwen2.5-7B-Instruct is a robust, multipurpose LLM suitable for financial Q&A, multilingual chat, structured outputs, and multi-turn reasoning, provided prompting and RAG integration are implemented externally. For strict guardrails or function execution, additional layers are required.